Distinction between phyllodes tumor and fibroadenoma in breast ultrasound using deep learning image analysis

Stoffel, Elina ; Becker, Anton S ; Wurnig, Moritz C ; Marcon, Magda ; Ghafoor, Soleen ; Berger, Nicole ; Boss, Andreas

Abstract: Purpose To evaluate the accuracy of a deep learning software (DLS) in the discrimination between phyllodes tumors (PT) and fibroadenomas (FA). Methods In this IRB-approved, retrospective, single-center study, we collected all ultrasound images of histologically secured PT (n = 11, 36 images) and a random control group with FA (n = 15, 50 images). The images were analyzed with a DLS designed for industrial grade image analysis, with 33 images withheld from training for validation purposes. The lesions were also interpreted by four radiologists. Diagnostic performance was assessed by the area under the receiver operating characteristic curve (AUC). Sensitivity, specificity, negative and positive predictive values were calculated at the optimal cut-off (Youden Index). Results The DLS was able to differentiate between PT and FA with good diagnostic accuracy (AUC = 0.73) and high negative predictive value (NPV = 100%). Radiologists showed comparable accuracy (AUC 0.60-0.77) at lower NPV (64-80%). When performing the readout together with the DLS recommendation, the radiologist's accuracy showed a non-significant tendency to improve (AUC 0.75-0.87, p = 0.07). Conclusion Deep learning based image analysis may be able to exclude PT with a high negative predictive value. Integration into the clinical workflow may enable radiologists to more confidently exclude PT, thereby reducing the number of unnecessary biopsies.

DOI: https://doi.org/10.1016/j.ejro.2018.09.002

European Journal of Radiology Open 5 (2018) 165–170

Contents lists available at ScienceDirect

European Journal of Radiology Open

journal homepage: www.elsevier.com/locate/ejro

Distinction between phyllodes tumor and fibroadenoma in breast ultrasound using deep learning image analysis

Elina Stoffel*, Anton S. Becker, Moritz C. Wurnig, Magda Marcon, Soleen Ghafoor, Nicole Berger, Andreas Boss

Institute for Diagnostic and Interventional Radiology, University Hospital of Zurich, Switzerland

ARTICLE INFO

Keywords:
Breast imaging
Ultrasound
Phyllodes
Fibroadenoma
Deep learning
Computer assisted diagnosis

ABSTRACT

Purpose: To evaluate the accuracy of a deep learning software (DLS) in the discrimination between phyllodes tumors (PT) and fibroadenomas (FA).

Methods: In this IRB-approved, retrospective, single-center study, we collected all ultrasound images of histologically secured PT (n = 11, 36 images) and a random control group with FA (n = 15, 50 images). The images were analyzed with a DLS designed for industrial grade image analysis, with 33 images withheld from training for validation purposes. The lesions were also interpreted by four radiologists. Diagnostic performance was assessed by the area under the receiver operating characteristic curve (AUC). Sensitivity, specificity, negative and positive predictive values were calculated at the optimal cut-off (Youden Index).

Results: The DLS was able to differentiate between PT and FA with good diagnostic accuracy (AUC = 0.73) and high negative predictive value (NPV = 100%). Radiologists showed comparable accuracy (AUC 0.60–0.77) at lower NPV (64–80%). When performing the readout together with the DLS recommendation, the radiologist's accuracy showed a non-significant tendency to improve (AUC 0.75–0.87, p = 0.07).

Conclusion: Deep learning based image analysis may be able to exclude PT with a high negative predictive value. Integration into the clinical workflow may enable radiologists to more confidently exclude PT, thereby reducing the number of unnecessary biopsies.

1. Introduction

Phyllodes tumor (PT) of the breast are rare breast lesions, accounting for less than 1% of all breast tumors. They are typically seen in women aged 35 to 55 years at presentation and are mostly large with a median size of 4 cm [1]. Histologically, they are characterized by "leaf-like" lobulations, from which the name is derived (Greek *phullon* leaf), with more abundant and cellular stroma than that of fibroadenoma (FA). PT are commonly classified into categories of benign, borderline, or malignant on the basis of histological parameters such as mitotic count, cellular atypia, stromal cellularity and overgrowth, and the nature of tumor borders [2]. Histologically, benign PT can be mistaken for FA, whereas at the other end of the spectrum, malignant PT show overlapping features with primary breast sarcomas or spindle cell metaplastic carcinoma. However, regardless of their histology, all PT can recur, where an increased risk of local recurrence is correlated with larger size and malignancy [3–5].

FA is the most common benign tumor of the breast in women under 35 years of age. They present as well-defined, mobile masses that can increase in size and tenderness in response to high levels of estrogen (e.g. during pregnancy or prior to menstruation). Histologically, they are made up of both glandular breast tissue and stromal tissue. In contrast to PT, risk of cancer is usually not increased in FA [6].

In addition to their histopathological similarities, FA are usually indistinguishable from PT on a macroscopic level. Both fibroepithelial tumors are often detected as fast growing breast lumps, and distinguishing PT from FA by means of physical exam is extremely difficult. With increased public awareness and screening, most of the breast tumors are being discovered at earlier stages, when both tumors share a substantial overlap in sonographic features and size [7,8]. Furthermore, sonography cannot distinguish between malignant, borderline and benign PT. Diagnostic evaluation is therefore often extended to the use of invasive diagnostic procedures, such as core-needle biopsies. However, even with the help of histology, diagnosis can be complicated due to sampling errors.

The diagnosis has wider implications that also influence the therapeutic approach to these tumors. Although conservative management is an acceptable strategy in FA, malignant PT should be completely

* Corresponding author at: Institute for Diagnostic and Interventional Radiology, University Hospital Zurich, Raemistrasse 100, 8091 Zurich, Switzerland.
E-mail address: Elina.stoffel@uzh.ch (E. Stoffel).

https://doi.org/10.1016/j.ejro.2018.09.002
Received 13 March 2018; Received in revised form 7 September 2018; Accepted 7 September 2018

enucleated with clear margins due to the high recurrence rate of up to 30% [9] with metastases and death being observed in 22% [2]. Without re-excision, the recurrence rates can be as high as 43%, necessitating an additional operation [10].

Traditionally, patients with breast masses that cannot clearly be identified as FA or PT will usually undergo complete surgical excision or mastectomy, for the fear of overlooking a potentially malignant tumor. Therefore, accurate identification and differentiation of PT preoperatively is critical to appropriate surgical planning, avoiding operative complications resulting from inadequate excision or surgical overtreatment. Most FA do not need surgical treatment at all. In these cases, biopsies are essentially an unnecessary physical, psychological and financial burden for the patient [11].

Deep learning is a type of machine learning that was inspired by the structure and function of the brain. It imitates the mammalian visual cortex in processing data using artificial neural networks (ANNs) that contain hidden layers. The deep learning software (DLS) learns to extract meaningful features from images to then make inferences and decisions on its own. "Meaningful" in this context stands for "helping to solve the problem at hand", in our case discriminating FA from PT. This data-driven method has shown promising results in recent years, as opposed to older more algorithmic approaches with hand-crafted features, which may often yield many arbitrary features not useful for the problem at hand. Hence, the use of deep learning in radiology as a method of differentiating and diagnosing tumors is a rapidly growing field [12]. Although, as with any diagnostic test, false-positive results can occur, the sensitivity of deep learning e.g. in mammography has reached numbers of up to 84%, equaling or surpassing the diagnostic accuracy of seasoned specialists [13]. Deep learning can be integrated into the assessment of sonographically detectable lesions and could be performed in the initial evaluation of indeterminate breast tumors (illustrated in Fig. 1).

In this retrospective, single-center study, we aimed to evaluate the precision of a DLS in the discrimination between PT and FA.

2. Materials and methods

2.1. Ultrasound examination and reference standard

This retrospective study was approved by the IRB, who waived the need for informed consent. All patients from a two-year period (July 2013 – July 2015) were reviewed for the presence of PT with histology as a reference standard (n = 11). From the remaining patients, a random subset with histologically secured diagnosis of a FA was taken (n = 15). Due to the low number (n = 4), FA with histopathological phyllodes features were counted towards one of the other groups. Since the management at our institution for those lesions is surgical excision, they were counted as PT. Median lesion diameter (long axis) was 21.5 mm (interquartile range 18–26 mm) for FA and 26.0 mm for PT (19–37 mm, p = 0.25). Lesion volume as calculated with all three diameters and the ellipsoid formula was also not significantly different (13.6 vs. 24.3 cm^3, p = 0.55). Mean age ± 95% confidence interval was 33.6 ± 15.2 years. All examinations were performed on the same type of ultrasound device (Logiq E9, GE Healthcare, Chicago, IL, USA) with the same reconstruction setting ("Breast"). For large lesions, multiple focus points were used. Functional ultrasound images were not consistently acquired and hence not included for analysis (i.e. with doppler or elastography overlay). For lesions depicted in multiple images, all available data was used, resulting in a total of 50 images of FA and 36 images of PT. The raw DICOM images were converted into lossless,

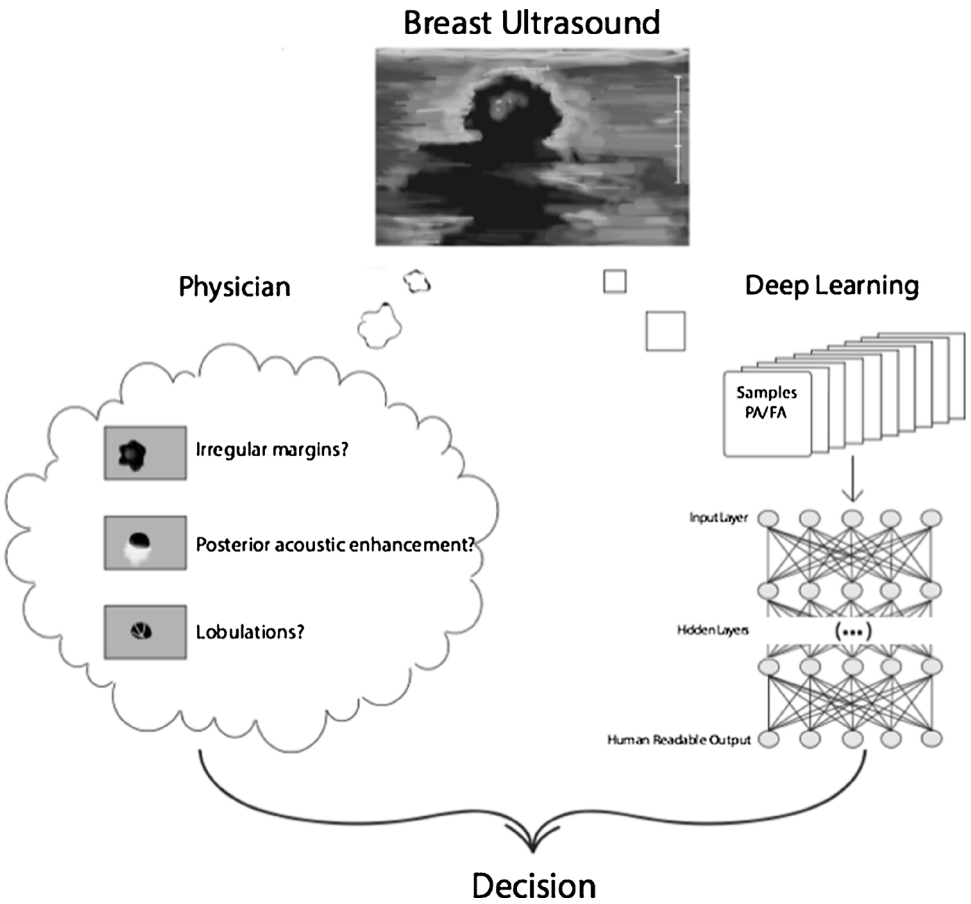

Fig. 1. Proposed integration of a deep learning based software into the clinical workflow. Deep learning image analysis has the ability to evaluate features, which are not perceptible to the human reader and may thus augment the evaluation of the radiologist.

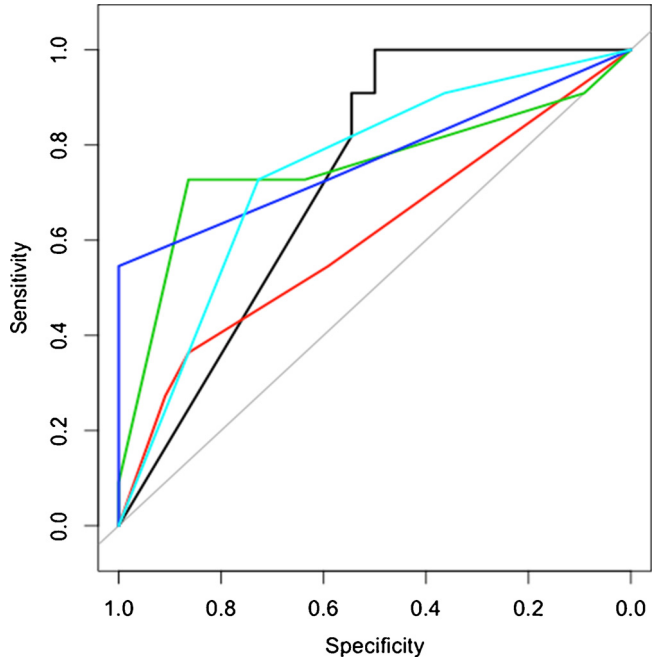

Fig. 2. When comparing the performance of human readers and DLS on the validation data set, the complimentary performance of the DLS with high sensitivity ($p < 0.05$ for all readers) but lower specificity is already evident (AUC reader 1 [red] = 0.60, reader 2 [green] = 0.77, reader 3 [dark blue] = 0.75 and reader 4 [light blue] = 0.74). Readers 1 to 4 represent radiologists from highest to lowest levels of experience, respectively.

monotone jpeg for further processing.

2.2. Deep learning image analysis

Image analysis was performed with a DLS originally developed for industrial image analysis (ViDi Suite Version 2.0; Cognex Inc, Natick MA, USA). This software takes advantage of the latest advances in deep learning algorithms to classify anomalies in images [14–16]. It is currently used in various industries for real-time quality inspection e.g. in defect detection of metal surfaces, traffic analysis or appearance based product identification. It is currently not FDA approved but has recently shown promising results for detecting cancer in a dual-center mammography study [17]. The exact architecture of the deep networks in ViDi Suite is proprietary, however, a comprehensive review about the broader topic can be found in [16]. All computations were performed on a GeForce GTX 1080 graphics processor unit. In a first step, the images were cropped to the actual lesion by using the supervised ViDi Detection Tool, the architecture of which is optimized for anomaly localization in homogeneous patterns (i.e. subcutaneous fat). In a second step, the cropped lesions were analyzed using the ViDi Classification Tool, which in turn is optimized for image classification. A randomly chosen subset of images (n = 53) was used for the training of the software (training set), and the remaining images (n = 33) were withheld from the software and solely used to validate the resulting model after training (validation set). The split was performed on a per-patient basis.

Table 2
Pairwise interreader agreement measured by weighted Cohen's Kappa (95%-CI in brackets).

	Reader 1	Reader 2	Reader 3	Reader 4
R1	1	0.3 (0.05–0.56)	0.21 (−0.01 to 0.43)	0.21 (−0.03 to 0.44)
R2	0.3 (0.05–0.56)	1	0.47 (0.22–0.71)	0.31 (0.07–0.55)
R3	0.21 (−0.01 to 0.43)	0.47 (0.22–0.71)	1	0.36 (0.14–0.58)
R4	0.21 (−0.03 to 0.44)	0.31 (0.07–0.55)	0.36 (0.14–0.58)	1

Table 3
Confusion matrices of the reader's performances.

		All		Valid.	
	Reference	FA	PH	FA	PH
Reader 1	FA	46	19	19	7
	PH	4	17	3	4
Reader 2	FA	32	10	13	3
	PH	18	26	9	8
Reader 3	FA	33	8	14	3
	PH	17	28	8	8
Reader 4	FA	22	12	8	1
	PH	28	24	14	10

2.3. Radiologist's readout

The cropped lesions were exported and presented to four radiologists in random order (INITIALS BLINDED FOR REVIEW: Board certified radiologists with 8, 3 and 2 years of experience in breast imaging, as well as a PGY-3 resident, referred to as reader 1–4, respectively). The readers were blinded to the study design as well as the clinical information of the patients. The images were rated on a 5-point Likert-like scale reflecting the confidence of the reader in his or her diagnosis (1 = definitely FA, 5 = definitely PT). After a four-month waiting period to avoid memory bias, the radiologists rated the lesions of the validation set again. This time, the DLS rating was shown below the image and the radiologists were asked to take it into consideration as well.

2.4. Statistical analysis

The statistical analysis was performed in R version 3.3.1 (R Foundation for Statistical Computing, Vienna, Austria). Continuous variables were expressed as median and interquartile range, categorical variables as counts or percentages. Interreader agreement was assessed pair-wise with the weighted Cohen's Kappa [18]. Kappa values (κ) were interpreted after the suggestion of Altman [19]: < 0.20, poor,

Table 1
Diagnostic performance (area under the ROC curve + 95% CI) and sensitivity, specificity, positive and negative predictive value (NPV) of the DLS for all cases as well as the validation set separately, and each reader (all data).

	AUC	Sensitivity	Specificity	NPV	PPV
DLS (All)	0.89 (0.83–0.95)	0.78 (.66–.9)	1	1	0.77 (.68–.88)
DLS (Valid.)	0.73 (0.59–0.87)	0.5 (.27–.72)	1	1	0.5 (.41–.65)
Reader 1	0.73 (0.62–0.84)	0.47 (.31–.64)	0.92 (.84–.98)	0.71 (.64–.78)	0.81 (.65–.95)
Reader 2	0.77 (0.68–0.87)	0.78 (.64–.91)	0.66 (.52–.78)	0.8 (.71–.91)	0.62 (.53–.73)
Reader 3	0.73 (0.63–0.83)	0.72 (.58–.86)	0.64 (.52–.76)	0.76 (.67–.86)	0.59 (.50–.70)
Reader 4	0.6 (0.48–0.71)	0.67 (.52–.80)	0.44 (.30–.81)	0.64 (.51–.77)	0.46 (.38–.54)

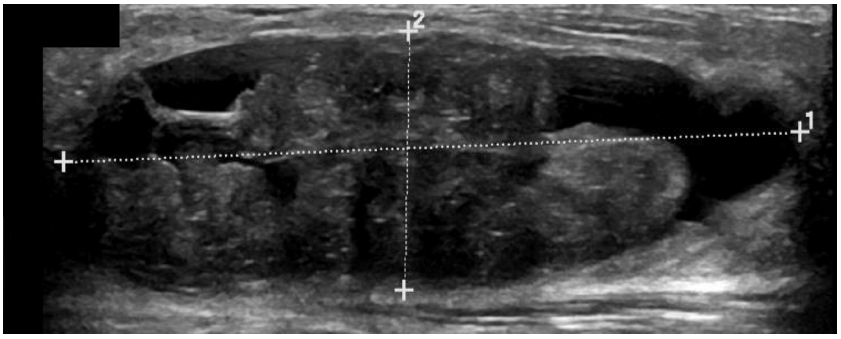

Fig. 3. Example of a true positive in both the DLS and readers. This large lesion exhibited irregular internal structure with cystic components and indistinct margins, which lead to the correct diagnosis of this PT by both the DLS and all 4 readers.

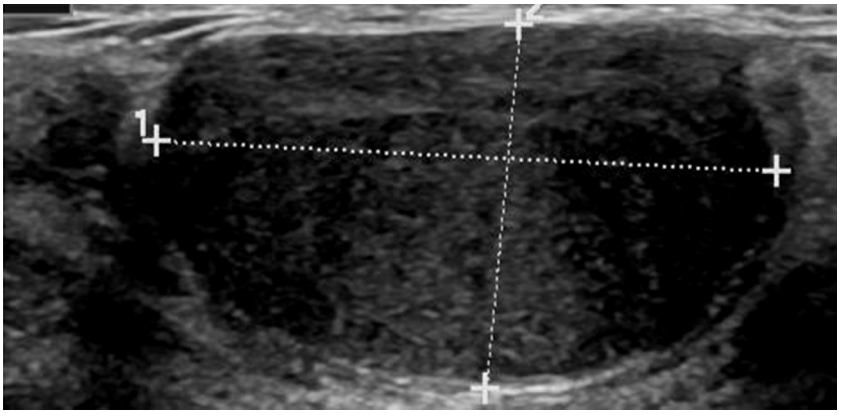

Fig. 4. Example of a false positive of the DLS and a true negative from experienced readers. The ultrasound image depicts a FA that the DLS falsely interpreted as a PT, possibly due to the acoustic shadowing at the borders. Three readers identified this image as a probable, and one reader as a definite FA.

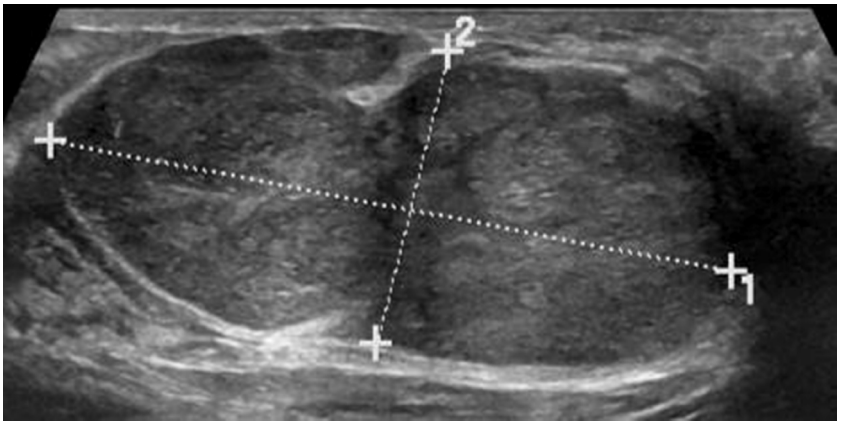

Fig. 5. Example of a true negative from the DLS and false positive from the readers. This FA from the validation set (i.e. withheld during training of the DLS) was correctly identified by the DLS, however, the readers falsely interpreted it as a probable PT (and one reader rated it as indeterminate).

0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80, good and 0.81–1.0 very good agreement.

Diagnostic performance was assessed with a receiver operating characteristic (ROC) analysis for the computer test and the human readers. Diagnostic accuracy was expressed as the area under the receiver operating characteristic curve (AUC) and compared with DeLong's nonparametric test for paired data [20]. Sensitivity, specificity, positive and negative predictive values were calculated at the optimal cut-off (Youden-Index). A p-value < 0.05 was considered indicative of significant differences. All tests were two-tailed.

3. Results

3.1. Deep learning image analysis

The DLS showed an excellent AUC of 0.89 on the whole data set, and an AUC of 0.73 on the validation data, indicating some overfit to be present in our rather small data set. In order to demonstrate the generalizability to new cases, Fig. 2 depicts the performance on the validation data only. On both training and validation data the DLS exhibited a high specificity of 1.0 (summarized in Table 1).

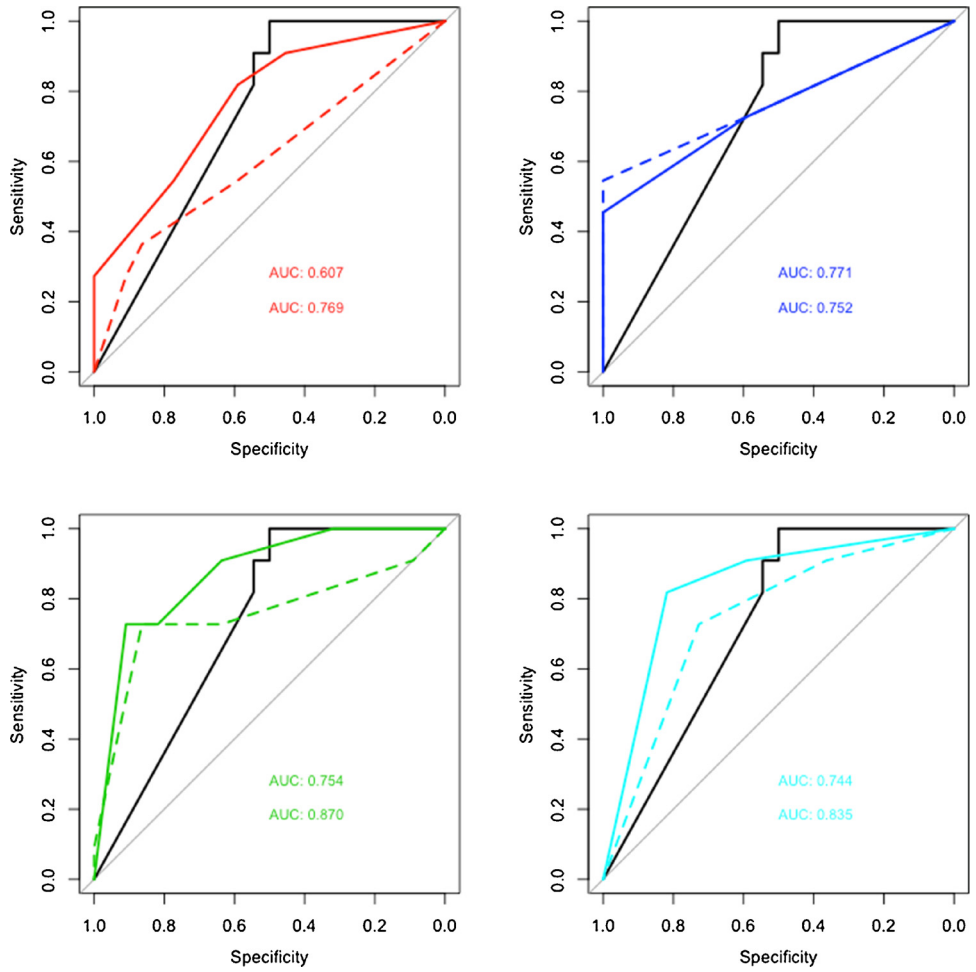

Fig. 6. ROC curves of Readers 1–4 (top left to bottom right) showing a non-significant tendency of improved performance in three out of the four readers (p_{min} = 0.07) when taking the DLS prediction into consideration (solid line), compared to the radiologists' judgement alone (dashed line).

3.2. Radiologist readout

Table 2 demonstrates that overall, interreader agreement was rated as fair (0.21–0.40) or moderate (0.41–0.60). Reader 1 showed the lowest interreader agreement (0.3, 0.21 and 0.21, respectively). Reader 2 and 3 exhibited the highest interreader agreement of 0.47 (moderate). The level of agreement for reader 4 gradually increases for each reader (0.21 for reader 1, 0.31 for reader 2 and 0.36 for reader 3). Readers 2, 3, and 4 show very similar AUCs (0.77, 0.75, and 0.74), while reader 1 showed the lowest AUC of 0.60. In the validation set (Fig. 2), reader 4 exhibited highest specificity and reader 3 the highest sensitivity. The confusion matrices with the readers' ratings vs. reference standard are summarized in Table 3.

3.3. Comparison of diagnostic performance

Diagnostic performance (AUC) was not significantly different between the DLS (validation data) and any of the readers (p-values 0.31, 0.80, 0.66, 0.87; example case shown in Fig. 3). However, at the optimal cut-off, the DLS was more sensitive than the readers, resulting in some false positives (example shown in Fig. 4) but consistently exhibited higher NPV and specificity (example case shown in Fig. 5).

In the second readout, three out of the four readers showed a non-significant tendency of improved performance (p = 0.07), with reader 1 improving the most (from 0.61 to 0.77), and reader 2 slightly decreasing (from 0.77 to 0.75). Readers 3 and 4 moderately improved (0.75 to 0.87 and 0.74 to 0.84). In general, there was a higher gain in specificity than in sensitivity as can be seen in the ROC curves in Fig. 6.

4. Discussion

In this pilot study, we have investigated whether a DLS can extract meaningful features from ultrasound image data and learn to distinguish PT from FA. We found that the software may be able to exclude PT with a high negative predictive value. Furthermore, we were able to show that combining the DLS estimate with the radiologist's impression leads to significantly better diagnostic performance.

The most widespread diagnostic and screening management of breast masses include physical examination, radiographic assessment (ultrasonography or mammography), and, if indicated, tissue specimen analysis (fine-needle aspiration or core needle biopsy). However, these diagnostic tests often fall short in differentiating PT from FA. Our results reflect the current controversy among radiologists in diagnosing PT and FA based on ultrasound images, evident by the poor interreader agreement in Table 1. Although MRI findings can be used to help determine the histological grade of known breast PT, MRI findings have been reported to be insufficient for reliable differentiation between FA and PT [21–23]. In this study, we show that deep learning image analysis can use ultrasound images to discriminate PT from FA with a specificity and negative predictive value that surpasses that of experienced radiologists. Furthermore, the software reached a diagnostic performance of 0.73 in the validation set, with the readers reaching comparable performance. Interestingly, the diagnostic performance of the radiologists did not correlate with their years of experience,

illustrating the ambiguity and lack of distinctive characteristics for either tumor. Furthermore, this could be due to the fact that the incidence of PT is very low and the radiologists had not been exposed to a high absolute number of cases despite years of experience (Table 2). Inter-reader agreement decreases for each reader, which may be explained by the gradually decreasing level of experience between readers (8, 3 and 2 years, and PGY-3 resident, respectively). Therefore, it is not surprising that reader 1 and 4 - the most experienced and the most inexperienced reader - show the lowest interreader agreement, and that readers 2 and 3 - with only 1 year difference training - demonstrate the most similar results.

When compared to DLS, the results show that the radiologists achieve higher readout specificity and thus positive predictive value - the ability to correctly identify PT - whereas deep learning image analysis showed the strongest performance in its negative predictive value - the ability to correctly exclude patients without PT. Hence, augmenting the reader's impression with the DLS estimate led to a significant increase in diagnostic accuracy. This improvement indicates that supplementation of deep learning image analysis into the diagnostic workup can enhance the accuracy in differentiating PT from FA. DLS is already integrated into routine diagnostics with a level of competence comparable to radiologists [17]. These results are in line with other fields outside of medical imaging: In chess, the combination of computer and amateur chess player outperforms either computer or chess grand master alone [24].

One of the limitations of our study design was that we only trained the software to distinguish between two different types of breast masses, PT and FA. This means that it cannot detect other lesions, such as invasive cancers or scars that may be important differential diagnoses. Furthermore, the software in its current form showed a high specificity and negative predictive value, meaning that it would correctly identify unaffected patients but not reliably identify patients who need treatment. This shortcoming seemed to be offset to a certain degree by using the software as a supplement to the radiologist's decision. Therefore, the momentary software would mostly be suitable as an adjunct tool to supplement a radiologist's diagnosis. Future studies and refinements of the software might allow deep learning to act as a screening tool for all types of breast lesions.

Further limitations of our study are the small sample size, the retrospective design as well as the restricted experimental setting. In the clinical routine, FA are far more common than PT. Since the software was trained on a cohort with a high prevalence of PT, it would possibly overestimate the occurrence of PT in the clinical routine. However, the high NPV should theoretically prevail or even increase.

DLS is novel method that has not yet been approved by the FDA or any other regulatory body. Furthermore, the cost-effectiveness of a DLS implementation has not yet been examined. These are some of the many questions that must be addressed before its broader use.

In conclusion, computer-assisted diagnosis in the form of deep learning image analysis is a useful tool to differentiate patients with PT and FA. A decision by the examining radiologist supplemented by the aid of DLS provides the highest diagnostic performance, and its integration into clinical routine may enable doctors to more confidently exclude PT, resulting in less unnecessary biopsies.

Funding/Disclosures

None.

Conflicts of interest

The authors of this manuscript declare no relevant conflicts of interest, and no relationships with any companies, whose products or services may be related to the subject matter of the article.

References

[1] S.P. Mishra, S.K. Tiwary, M. Mishra, A.K. Khanna, Phyllodes tumor of breast: a review article, ISRN Surg. 2013 (2013).
[2] S.R. Lakhani, I.O. Ellis, S.J. Schnitt, P.H. Tan, M.J. van de Vijver, WHO Classification of Tumours, Fourth ed., IARC WHO Classification of Tumours, 2012.
[3] R.J. Barth Jr., Histologic features predict local recurrence after breast conserving therapy of phyllodes tumors, Breast Cancer Res. Treat. 57 (1999) 291–295.
[4] G.M. Tse, P.C. Lui, C.S. Lee, F.Y. Kung, R.A. Scolyer, B.K. Law, T.S. Lau, R. Karim, T.C. Putti, Stromal expression of vascular endothelial growth factor correlates with tumor grade and microvessel density in mammary phyllodes tumors: a multicenter study of 185 cases, Hum. Pathol. 35 (2004) 1053–1057.
[5] R.Z. Karim, S.K. Gerega, Y.H. Yang, A. Spillane, H. Carmalt, R.A. Scolyer, C.S. Lee, Phyllodes tumours of the breast: a clinicopathological analysis of 65 cases from a single institution, Breast 18 (2009) 165–170.
[6] V.D. Yagnik, Juvenile giant fibroadenoma, Clin. Pract. 1 (2011) 98–99.
[7] T.C. Chao, Y.F. Lo, S.C. Chen, M.F. Chen, Sonographic features of phyllodes tumors of the breast, Ultrasound Obstet. Gynecol. 20 (2002) 64–71.
[8] I.K. Komenaka, M. El-Tamer, E. Pile-Spellman, H. Hibshoosh, Core needle biopsy as a diagnostic tool to differentiate phyllodes tumor from fibroadenoma, Arch. Surg. 138 (2003) 987–990.
[9] P.H. Tan, A.A. Thike, W.J. Tan, M.M. Thu, I. Busmanis, H. Li, W.Y. Chay, M.H. Tan, Phyllodes Tumour Network Singapore, predicting clinical behaviour of breast phyllodes tumours: a nomogram based on histological criteria and surgical margins, J. Clin. Pathol. 65 (2012) 69–76.
[10] M.L. Hopkins, T.S. McGowan, G. Rawlings, F.F. Liu, A.W. Fyles, J.L. Yeoh, L. Manchul, W. Levin, Phylloides tumor of the breast: a report of 14 cases, J. Surg. Oncol. 56 (1994) 108–112.
[11] H. Tan, S. Zhang, H. Liu, W. Peng, R. Li, Y. Gu, X. Wang, J. Mao, X. Shen, Imaging findings in phyllodes tumors of the breast, Eur. J. Radiol. 81 (2012) 62–69.
[12] D. Shen, G. Wu, H. Suk, Deep learning in medical image analysis, Annu. Rev. Biomed. Eng. 19 (2017) 221–248.
[13] A.S. Becker, M. Mueller, E. Stoffel, M. Marcon, S. Ghafoor, A. Boss, Classification of breast cancer from ultrasound imaging using a generic deep learning analsis software: a pilot study, Br. J. Radiol. 91 (Feb. (1083)) (2018) 20170576, https://doi.org/10.1259/bjr.20170576 Epub 2018 Jan 10.
[14] J. Sun, R. Wyss, A. Steinecker, P. Glocker, Automated fault detection using deep belief networks for the quality inspection of electromotors, Tm-Technisches Messen 81 (2014).
[15] G.E. Hinton, S. Osindero, Y.W. Teh, A fast learning algorithm for deep belief nets, Neural Comput. 18 (2006) 1527–1554.
[16] Y.A. LeCun, L. Bottou, G.B. Orr, K. Müller, Efficient BackProp, G. Orr, K. Müller (Eds.), Neural Networks: Tricks of the Trade, Springer-Verlag, Berlin, Heidelberg, 2012, pp. 9–48.
[17] A.S. Becker, M. Marcon, S. Ghafoor, M.C. Wurnig, T. Frauenfelder, A. Boss, Deep learning in mammography: diagnostic accuracy of a multipurpose image analysis software in the detection of breast cancer, Invest. Radiol. 52 (2017) 434–440.
[18] J. Cohen, Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit, Psychol. Bull. 70 (1968) 213–220.
[19] D.G. Altman, Practical Statistics for Medical Research, First edition, Chapman and Hall, Oxford, 1991.
[20] E.R. Delong, D.M. Delong, D.L. Clarke-Pearson, Comparing the areas under two or more correlated receiver operating characterstic curves: a nonparametric approach, Biometrics 44 (1988) 837–845.
[21] D.M. Farria, D.P. Gorczyca, S.H. Barsky, S. Sinha, L.W. Bassett, Benign phyllodes tumor of the breast: MR imaging features, AJR Am. J. Roentgenol. 167 (1996) 187–189.
[22] K.K. Kinoshita, T. Fukutomi, Magnetic resonance imaging of benign phyllodes tumors of the breast, Breast J. 10 (2004) 232–236.
[23] S. Wurdinger, A.B. Herzog, D.R. Fischer, C. Marx, G. Raabe, A. Schneider, W.A. Kaiser, Differentiation of phyllodes breast tumors from fibroadenomas on MRI, AJR Am. J. Roentgenol. 185 (2005) 1317–1321.
[24] G. Kasparov, Deep Thinking: Where Machine Intelligence Ends and Human Creativity Begins, First ed., PublicAffairs, 2017.